Т.О. Шевина

ОСНОВНЫЕ ПРОБЛЕМЫ АВТОМАТИЗАЦИИ ПОЛИТИЧЕСКОГО КОНТЕНТ-АНАЛИЗА СМИ И ПУТИ ИХ РЕШЕНИЯ

(ФГБОУ ВПО «Самарский государственный аэрокосмический университет им. академика С.П. Королева (национальный исследовательский университет)»)

Одной из проблем современных информационных систем является проблема автоматизированной обработки материалов, представленных на естественном языке. Особенно остро данная проблема проявляется при контент-анализе текстов, поскольку данный вид анализа опирается на содержание текста. Контент-анализ – исследовательская техника для получения выводов путем анализа содержания текста о состояниях и свойствах социальной действительности [6]. В рамках данной работы решается задача оценки влияния сообщений СМИ на рейтинг политической фигуры или партии.

Контент-анализ включает 3 основных этапа:

1) выделение единиц анализа, которые затем сводятся в категории анализа и переводятся в машиночитаемый вид;
2) подсчет частот категорий, при этом применяются различные инструменты математического аппарата для выявления взаимосвязей между ними;
3) интерпретация полученных результатов

При автоматизации контент-анализа выделяются проблемы, при решении которых необходимо участие эксперта:

- оценка межкодировочной надежности на этапе подготовки данных;
- выделение ключевых слов;
- построение понятийных структур;
- интерпретация результатов исследования.

Процесс решения подобных проблем невозможно запрограммировать стандартными средствами, поэтому возникает необходимость использования специальных алгоритмов обработки.

На этапе подготовки обучающих данных необходимо измерить достоверность поступающих данных, так как в случае недостоверных данных, дальнейшее проведение анализа бессмысленно. Под *надежностью (достоверностью)* в классическом контент-анализе понимается высокая вероятность того, что два и более кодировщика, работая независимо друг от друга, присвоили одни и те же коды одним и тем же элементам в каждом из проанализированных текстов. Данный тип надежности, называемый *межкодировочной надежностью*, может вычисляться различными методами. Низкий уровень вычисленной надежности свидетельствует о том, что полученным данным доверять нельзя [1].

Холсти для вычисления межкодировочной надежности для номинальных категорий рекомендует крайне простую *формулу* [3]:

$$R = \frac{2 * M}{N_1 + N_2},$$

где R – показатель надежности (Reliability),
M – общее количество совпавших кодовых элементов у двух кодировщиков,
$N_1 + N_2$ – общее количество закодированных элементов у 1-го и 2-го кодировщика соответственно.

Однако если категория имеет два значения (да — нет), то при кодировке двух выборок с помощью генератора случайных чисел надежность будет приблизительно равна 50%, что, очевидно, неверно.

Перро и Лей разработали другой подход к коррекции межкодировочной надежности [4]. Эта формула также не свободна от недостатков, в частности, она зависит от количества размерностей в категории. В каноническом виде она выглядит следующим образом:

$$I = \sqrt{(\frac{F_0}{N} - \frac{1}{k}) * \frac{k}{k-1}},$$

где F_0 – количество совпавших кодовых элементов,
N – общее количество кодовых элементов,
k – количество размерностей в категории[3].

Одной из наиболее часто применяемых мер согласия является «*каппа Коэна*», вычисляемая по формуле [2]:

$$k = \frac{P_a - P_в}{1 - P_в},$$

где

$$P_a = \sum_{i=1}^{j} p_{ij},$$

$$P_в = \sum_{i=1}^{j} (p_j + p_{+i}),$$

p_{ij} – вероятность того, что кодировщик 1 закодирует объект размерностью i, а кодировщик 2 закодирует этот же объект размерностью j;
p_{+i} – вероятность того, что второй кодировщик закодирует объект размерностью i,
p_j – вероятность того, что первый кодировщик закодирует объект размерностью j.

Значения «*каппы Коэна*» лежат на отрезке [-1;1], минус единице соответствует полное несогласие кодировщиков в оценке, 1 – полное согласие. Особый случай – когда каппа равна 0. Это означает, что с таким же успехом можно было бы попросить оценить объекты два генератора случайных чисел.

Следует отметить, что при поступлении данных в режиме обучения экспертной системы приемлемым считается уровень надежности не менее 80%.

Следующим важным этапом обработки данных является процесс выделения ключевых слов и их категоризация. Принцип выделения ключевых слов заключается в очищении текста от слов, не несущих семантической нагрузки и последующего стемминга оставшихся слов (в наиболее простом случае, стемминг – выделении неизменяемой части слова), затем расчете частотного рас-

пределения получившихся основ в тексте. Подробно эта проблема была рассмотрена в более ранних работах [7, 8].

Построение понятийных структур в контент-анализе необходимо для выделения области влияния отдельно взятого слова. *Понятийное поле* содержит ключевое слово со своим понятийным контекстом, который представлен в форме синонимов первого и второго круга. Определение понятийного поля состоит в формировании смысловых блоков текста из слов, исходя из их значения. При этом выбранные ключевые слова должны являться значимыми для данного текста. В терминах экспертных систем понятийной структуре соответствует понятие модели представления знаний из базы знаний.

Наиболее распространёнными моделями представления знаний являются логические модели, продукционные модели, фреймовые модели и семантические сети. Самыми удобными моделями для анализа смысловой нагрузки текста являются семантические сети.

Под *семантикой*, с точки зрения информационных технологий, можно понимать принципы организаций языковых конструкций естественного языка.

Под *семантической моделью* текста можно понимать эквивалент данного текста, представленный таким образом, чтобы анализ смысловой нагрузки текста мог быть выполнен с использованием автоматизированных систем.

Таким образом, семантическая сеть является наиболее удобной моделью представления знаний, поскольку заключает в себе логику построения понятийного поля слова (семантику).

Интерпретация результатов исследования реализуется средствами подсистемы объяснений экспертной системы.

Поскольку необходимо установить, позитивно или негативно влияет статья СМИ на рейтинг политика, при контент-анализе используются оценочные суждения. Для вычисления соотношения положительных и отрицательных относительно избранной позиции суждений используется формула коэффициента Яниса [5]. В случае, когда число положительных оценок превышает число отрицательных, коэффициент Яниса подсчитывается по формуле:

$$c = \frac{f^2 - f * n}{r * t},$$

где f – число положительных оценок; n – число отрицательных оценок;

r – объем содержания текста, имеющего прямое отношение к изучаемому признаку; t – общий объем анализируемого текста.

В случае, когда число положительных оценок меньше, чем отрицательных, коэффициент Яниса находится по формуле:

$$c = \frac{f * n - f^2}{r * t}.$$

Использование коэффициента Яниса считается более эффективным, так как учитывается, что при увеличении объема текста, относящегося к изучаемому признаку, но не содержащего оценок, соотношение должно уменьшаться.

Таким образом, в ходе исследования найдены следующие варианты решения проблем автоматизации, описанные в таблице 1.

Таблица 1. Варианты решения проблем

№	Проблема	Вариант решения
1	Оценка межкодировочной надежности	Оценка Холсти, оценка Перро-Лея, «каппа Коэна»
2	Выделение ключевых слов	Использование статистических методов поиска ключевых слов, использование лингвистических алгоритмов обработки слова
3	Построение понятийных структур	Использование знаний эксперта (средства экспертной системы)
4	Интерпретация результатов исследования	Расчет коэффициента Яниса, использование подсистемы объяснений экспертной системы

Литература

1. http://www.rek36.ru/otsenka-validnosti-dannyih.html (дата обращения: 25.09.2012)

2. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960

3. Holsti O. Content analysis for the social sciences and humanities. Reading, MA: Addison-Wesley Publishing Co., 1969

4. Pereauit W.D., Leigh L.E. Reliability of nominal data based on qualitative judgments // Journal of Marketing Research, 1989

5. Кутлалиев А., Попов А. Эффективность рекламы. – М.: Эксмо, 2006

6. Тихонов В.И. Проблемы категоризации при контент-анализе // Круг идей: модели и технологии исторической информатики: Труды III конференции Ассоциации «История и компьютер»

7. Шевина Т.О. Нейронечеткое прогнозирование рейтинга политических фигур на основе контент-анализа СМИ // V Международная науч.-практич. конф. учащихся и студентов: тезисы докладов конференции. – Протвино, 2012

8. Шевина Т.О. Решение задачи классификации заявок с помощью нейронечеткой продукционной сети Ванга-Менделя // Международная молодежная конф. «Королевские чтения»: тезисы докладов конференции. – Самара, 2011